AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology and Molecular Biology Reviews®

REVIEW

Computational Tools for the Analysis of Uncultivated Phage Genomes

Juan Sebastián Andrade-Martínez,[a] Laura Carolina Camelo Valera,[a] Luis Alberto Chica Cárdenas,[a] Laura Forero-Junco,[a,b] Gamaliel López-Leal,[a] J. Leonardo Moreno-Gallego,[a,c] Guillermo Rangel-Pineros,[a,d] Alejandro Reyes[a,e]

[a]Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia
[b]Department of Plant and Environmental Science, University of Copenhagen, Frederiksberg, Denmark
[c]Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany
[d]The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[e]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA

SUMMARY Over a century of bacteriophage research has uncovered a plethora of fundamental aspects of their biology, ecology, and evolution. Furthermore, the introduction of community-level studies through metagenomics has revealed unprecedented insights on the impact that phages have on a range of ecological and physiological processes. It was not until the introduction of viral metagenomics that we began to grasp the astonishing breadth of genetic diversity encompassed by phage genomes. Novel phage genomes have been reported from a diverse range of biomes at an increasing rate,

Address correspondence to Alejandro Reyes, a.reyes@uniandes.edu.co.

The authors declare no conflict of interest.

Published 21 March 2022

which has prompted the development of computational tools that support the multilevel characterization of these novel phages based solely on their genome sequences. The impact of these technologies has been so large that, together with MAGs (Metagenomic Assembled Genomes), we now have UViGs (Uncultivated Viral Genomes), which are now officially recognized by the International Committee for the Taxonomy of Viruses (ICTV), and new taxonomic groups can now be created based exclusively on genomic sequence information. Even though the available tools have immensely contributed to our knowledge of phage diversity and ecology, the ongoing surge in software programs makes it challenging to keep up with them and the purpose each one is designed for. Therefore, in this review, we describe a comprehensive set of currently available computational tools designed for the characterization of phage genome sequences, focusing on five specific analyses: (i) assembly and identification of phage and prophage sequences, (ii) phage genome annotation, (iii) phage taxonomic classification, (iv) phage-host interaction analysis, and (v) phage microdiversity.

KEYWORDS computational analysis, microdiversity, phage annotation, phage metagenomics, phage taxonomy, phage and prophage identification, phage-host interaction, uncultivated viruses, viromes

INTRODUCTION

The use of metagenomics to discover and characterize populations of microbes and viruses in a particular niche is increasingly common (1). High-throughput sequencing (HTS) has made it possible to identify new populations of microorganisms without the need for them to be cultured or dependent on the use of specialized isolation methods (1). The identification and analysis of viruses in environmental and microbiome settings are relevant for multiple fields, most notably human health (2).

With an abundance estimated at 10^{31} particles, viruses are the most numerous biological entities in the biosphere (3). Metagenomic analyses have greatly contributed to the elucidation of the true diversity of viruses: early studies revealed that while most of the cellular diversity in the biosphere had already been discovered by the mid-2000s, a good deal of the global virome remained either unknown or unclassified (4). This is especially true for bacteriophages (phages), viruses that infect bacteria. As of 2021, there are over 12,000 complete and taxonomically classified phage genomes deposited in NCBI databases (5); however, single metagenomic studies can potentially identify up to several thousand new and unclassified phage genomes (6, 7). Those which cannot be fully classified or annotated remain as so-called “viral dark matter” (8). Most of these phages cannot be cultured, with uncultivated viral genomes (UViGs) making up more than 95% of the current diversity in public databases (9), for which we usually ignore the morphology and host, which are traditionally the most relevant features employed for assigning phage taxonomy.

Given the differences in viral genomic sequences, both between distinct types of viruses and between viruses and cellular organisms, viral metagenomics poses multiple challenges (8). In consequence, a plethora of metagenomic tools and pipelines have been developed to handle every aspect of these analyses (8, 10), some of them specialized for use in phages and some designed for all viral data. This makes it difficult to keep up with the available offerings and to select a specific tool with adequate parameters for a specific set of metagenomic data.

In that context, this review discusses some of the available software for metagenomics analysis of phage data, focusing on five topics: (i) assembly and identification of phage and prophage sequences, (ii) phage genome annotation, (iii) phage taxonomic classification, (iv) phage-host interaction analysis, and (v) phage microdiversity (Box 1; Fig. 1). We classified the tools based on the purpose they were designed for, and we briefly point out the most relevant (in our opinion) factors to consider when selecting one over the others. Note that while some of the tools mentioned can be

BOX 1: GLOSSARY

- Area under the ROC curve (AUC): Statistical measurement of the performance of a classifier, based on sensitivity and specificity. Its values range from 0 to 1, with 1 being the best possible. While a value of 0 is considered the worst possible, any value under 1/*N*, with *N* being the number of classes which can be predicted, is worse than random.
- Command line: Also referred to as a terminal, an interface through which a user can interact with a computer or server without a graphical interface. Communication via the command line is done through the typing and execution of commands from a shell programming language (e.g., Bash).
- E value: number of sequences in a database which would be expected to have, by chance, an alignment as good as or better than the one obtained given a query sequence and the selected database. The lower the E value, the more significant the alignment, and the more likely that the aligned sequences are homologous.
- F1 score: Statistical measurement of the performance of a classifier, based on precision and recall. Its values range from 0 to 1, with 1 being the best value possible and 0 the worst.
- FASTA format: Format for the computational representation of nucleotide or amino acid sequences. A sequence in FASTA format is composed of (i) one description line, identified by a leading greater-than (>) sign, which contains the description of the sequence, and (ii) a sequence line, made up of the sequence itself without any additional characters.
- General feature format (GFF): Format for the storage of descriptions regarding biologically relevant features in a nucleotide or amino acid sequence. The GFF format is tab delimited, with one line per feature.
- Profile hidden Markov models (pHMMs): Mathematical representations of a set of conserved regions in a given group of sequences. pHMMs are derived from multiple sequence alignments and are particularly useful for searching for distantly related sequences.
- k-mer: Sequence of either nucleotides of amino acids of length k. For example, ATCG is a tetramer, or 4-mer, while LME is an amino acid trimer, or 3-mer.
- Microdiversity: Intrapopulation genetic variation.
- RefSeq: Secondary (i.e., curated) nucleotide and amino acid database managed by the National Center for Biotechnology Information (NCBI).
- Single nucleotide polymorphism (SNP): A type of mutation in which a single nucleotide is changed to another.
- t-distributed stochastic neighbor embedding (t-SNE): Statistical method employed for the graphic visualization of distance data.
- Threshold: In the context of computational biology, a specific value of a parameter in a tool, pipeline, or method, which is set as the maximum or minimum value admissible for a result to be called or considered significant.

used for eukaryotic viruses, and even for cellular organisms, we focus on applications for phage data. We do not aim at recommending any particular tool above others; instead, we want to make the reader aware of the different advantages and limitations of the most used, available tools and to outline the main factors to be considered when selecting a tool to make an informed decision now and as more tools are developed.

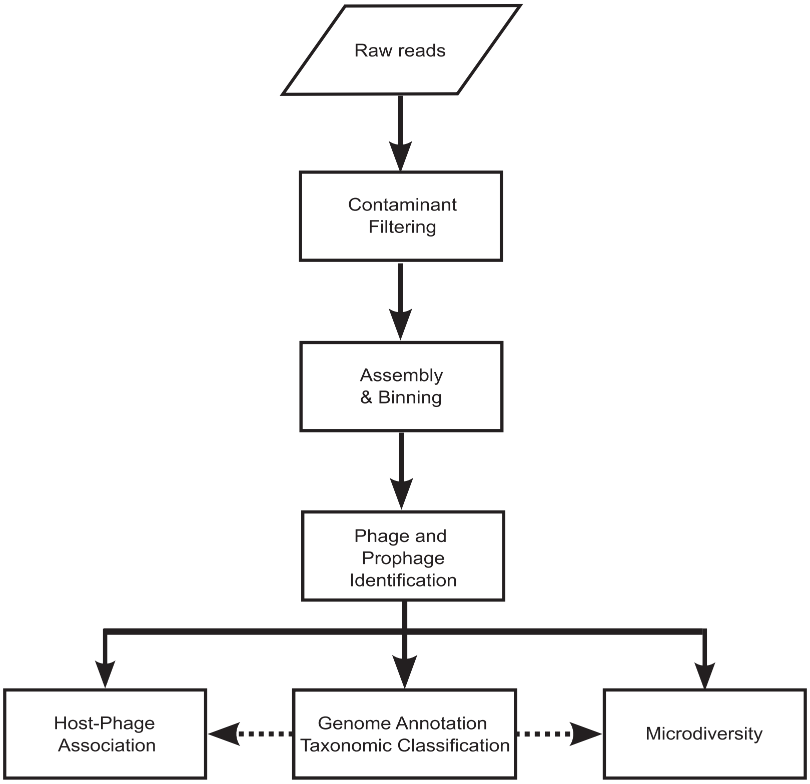

FIG 1 Proposed workflow for the analysis of phage metagenomic data. Raw metagenomic reads are first filtered for contaminants and then assembled into contigs and binned. While in principle the identification of phage and/or prophage contigs could be omitted if the researcher knows that the reads are enriched for viruses, we suggest performing it as an additional way to filter contaminant nonviral bins. Phage or prophage contigs can then be subjected to genome annotation, taxonomic classification, microdiversity analysis, and host-association analysis. Moreover, while they are different analyses, genome annotation and taxonomic classification are usually done conjointly. While one can carry out microdiversity and/or phage-host analysis without prior annotation and/or taxonomic classification, the former two are usually done before either of the latter.

ASSEMBLY AND IDENTIFICATION OF PHAGES AND PROPHAGES IN METAGENOMIC DATA

The process of phage community characterization from a metagenomic data set usually begins with the assembly and identification of phage genomic sequences (11). These sequences are largely derived from actively infecting phages or from prophages integrated into the genomes of their bacterial hosts (8). Thus, to thoroughly characterize the structure of phage communities from sequenced metagenomes, it is critical to have the ability to detect phage-derived sequences from either of the aforementioned sources and differentiate them from other potential contaminants or nonviral nucleic acids. Achieving an optimal balance between these two properties (recall and precision, respectively) is critical in order to obtain the most accurate picture of the phage community structure present within a metagenome (12). During the last two decades, several phage and prophage detection tools have been developed, and they encompass a wide range of strategies for detecting phages and/or prophages in metagenomic data. In general, all these tools take metagenomic assemblies as input data; thus, the assembly of sequencing reads is a critical step and marks the process where phage and prophage detection begins.

Assembly and Binning Approaches for Viral Metagenomic Data

Main assembly approaches. The selection of an assembly software for viral metagenomic data is critical for an accurate identification of viral contigs and other downstream analyses. The efficacy of the 16 most common short-read assemblers, specialized in metagenomic data or not, was recently thoroughly reviewed by Sutton et al.

(13) in three distinct data sets composed of both real and artificial viromes. Overall, the authors recommend the use of MetaSPAdes (14) for virome assembly, which yielded good results in all the test sets considered, followed by MEGAHIT (15). They also highlight the presence of repeat sequences, as well as too-high and too-low coverage values, as the main hindrances to efficient assembly of virome data. In particular, given that MetaSPAdes performed poorly in the assembly of poorly covered viral genomes, they suggest that the conjoint use of MetaSPAdes and MIRA (16) might be able to provide an overall better assembly than either program employed individually.

ViralAssembly (17), developed as part of the MetaviralSPAdes pipeline, is an adaptation of MetaSPAdes for assembly of viral data. It leverages the circular genomic sequence detection of MetaplasmidSPAdes for detecting circular viral genomes and allows detection and assembly of terminal repeats in linear genomes. In an analysis of 18 real virome data sets, ViralAssembly was shown to outperform MetaSPAdes in terms of contig completeness in 12 cases. Furthermore, rnaSPAdes (18), originally designed for the assembly of transcriptomic data, was recently shown to have the capacity to generate RNA phage contigs from metagenomic data (19).

One of the most exciting prospects in viral assembly is the development of assembly software for long-read sequencing technologies, which might allow not only increased viral detection but also more resolution in elucidating the microdiversity of viral communities (17, 20). While not a specific viral assembler, metaFlye (21), based on the generation of assembly graphs via high-frequency k-mers, has been shown to detect and assemble viral genomes in long-read metagenomic data sets with good efficiency. A specific version for the assembly of viral genomes, viralFlye, is currently in development (https://github.com/Dmitry-Antipov/viralFlye). The VirION pipeline, now in its second iteration, VirION2 (22), employs short reads to correct sequencing errors in long-read assemblies and outperforms hybrid and short-read assemblers when tested on double-stranded-DNA (dsDNA) viromes. It is important to mention that to the best of our knowledge, the efficacy of said assembly method has not been tested yet for RNA or single-stranded-DNA (ssDNA) viruses.

From the final set of the three recommended software programs, MetaviralSPAdes and metaFlye are available for download from GitHub, where the user can find instructions on how to install and run them. Both require a basic knowledge of command line usage for installation, as well as installation of certain dependencies. For the case of VirION2, the complete protocol, including its experimental steps, is available online (https://www.protocols.io/view/virion-2-6q9hdz6). Each individual bioinformatic tool must be installed and run by the user.

Selecting an Assembly Tool

The choice of a specific assembler will mostly depend on the type of reads. For short reads, we follow the recommendation of Sutton et al. (13), albeit using MetaviralSPAdes as opposed to MetaSPAdes, which can be combined with MIRA if low-abundance genomes are present. For long reads, a user with little computational experience might benefit from using metaFlye or viralFlye, while someone more experienced might be better off replicating the VirION2 pipeline, provided that sequencing was done via the adequate technology.

As a final note, it is recommended that the researchers remap all the unassembled reads, if any, to the contigs generated by their choice of assembly software. This process has been shown to improve downstream taxonomic classification (23) and should also help to reduce the number of unassigned reads prior to binning.

Main binning approaches. A common issue in metagenomic assemblies is the generation of fragmented or partial assemblies; this is mainly due to variation in coverage or repeats that will often break the assemblies. Following the assembly, the reconstruction of microbial genomes from metagenomic data sets relies on the ability to group the assembled contigs in a way that resembles the genomes they were derived from, a process more commonly known as binning. Machine learning techniques are the basis of some recent virus-specific binning software, aimed at clustering contigs coming from the same viral species in a metagenomic sample. For example, CoCoNet (24) employs tetranucleotide frequencies and read coverage to train a neural network which computes the probability that two fragments come from the same genome and then clusters contigs into bins based on said probabilities. PHAMB (25), in contrast, uses a random forest classifier to discern the viral bins generated via deep variational autoencoders with VAMB (26). METABAT (27), now in its second iteration, METABAT2 (28), uses as input tetranucleotide frequencies and read coverage information, just like CoCoNet, but derives distances between contigs from these measurements and uses them directly for the clustering into bins. Of these three tools, only PHAMB discerns viral from nonviral bins directly, and it therefore might be preferred for use with metagenome data, while CoCoNet is specialized for use in highly diverse viromes. METABAT is not specialized in viromes, but output bins from METABAT can be designated as viral via other phage and prophage software or the taxonomic classification tools discussed in depth below.

Selecting a Binning Tool

All three software programs need to be installed by the user and require basic command line knowledge. CoCoNet represents the simplest installation, although the use of METABAT via Docker is also straightforward. For users with no computational experience, we recommend the use of CoCoNet. In contrast, for users with basic or higher command line knowledge, PHAMB might be preferred.

Detection of Phages and Prophages

Main approaches for phage and prophage identification. In general, all currently available phage and prophage identification tools employ one the following approaches. Approach 1 is detection of phage proteins through the search of homologous sequences in phage-specific databases that consist of amino acid sequences or profile hidden Markov models (HMMs). This search is coupled with the use of sliding windows to identify regions enriched in phage genes and other phage-associated properties (e.g., gene coding density, enrichment in hypothetical genes, enrichment in short genes, and depletion in strand switch). Approach 2 is the use of either supervised or deep learning prediction models to identify phage contigs, based on the calculation of sequence features that are independent of database searches. Approach 3 is a hybrid method that employs machine learning models based on both database search-dependent and -independent features.

Table 1 provides a summary of some of the phage prediction tools that are currently publicly available. The initial step to identify bona fide phage proteins in approaches 1 and 3 is based on the analysis of the best hits obtained through BLAST, hmmscan, or other similarity-based methods. A critical aspect of all those methods is the selected database, as the breadth of reported matches is influenced not only by the composition of the database (i.e., which viruses are present and how diverse they are) but also by its size, since this is known to affect the E value. For example, VirSorter (29), VIRALVERIFY (17), PHASTER (30), and ProphET (31) use the RefSeq viral database from the NCBI or selected groups of phages from it, although VirSorter also uses a custom database that includes phages from a variety of viromes spanning different types of environments (further discussed in the following section). On the other hand, Phigaro (32) and VirMiner (11) perform HMM searches against the prokaryotic Virus Orthologous Groups (pVOGs) database, which is significantly smaller (33). It is important to note that most of these tools use similar thresholds. For example, PHASTER (30) and ProphET (31) consider significant (phage-like matches) hits with an E value of $<10^{-4}$, while VirSorter uses 10^{-5}. However, in order to provide higher confidence in the viral predictions reported to the user, these tools consider further virus-associated parameters, such as protein length, transcription strand directionality, customized AT and GC skew, phage insertion sites, and the identification of tRNA genes (29–31, 34).

Phage prediction methods that employ machine learning (approaches 2 and 3) calculate a set of features from the input nucleotide sequences, and these are then fed into a prediction model that determines whether the input sequences are likely to be

TABLE 1 Comparison of the different tools presented for identification of phage and prophage sequences

Tool	Type[a]	Input data type	Accessibility	Last update
VirSorter (29)	1	Viral or phage genomes or contigs; host genomes for prophage prediction (FASTA files)	Web-based (https://cyverse.org/) and stand-alone versions	Last release Oct 2019
VirSorter2 (36)	3	Viral or phage genomes or contigs; host genomes for prophage prediction (FASTA files)	Web-based (https://cyverse.org/) and stand-alone versions	Last update Apr 2021
VirFinder (37)	2	Viral or phage genomes or contigs	Stand-alone version	Last update Sept 2019
DeepVirFinder (42)	2	Viral or phage genomes or contigs	Stand-alone version	Last update Nov 2020
MARVEL (39)	3	Viral or phage genomes or contigs and raw reads	Stand-alone version	Last update Apr 2019
PPR-Meta (35)	2	Phage and plasmid fragments from metagenomic assemblies	Stand-alone version	Last update Jan 2020
VIBRANT (38)	3	Sequence derived from metagenomic assemblies	Web-based (https://cyverse.org/) and stand-alone versions	Last update May 2020
VirMiner (11)	3	Processed raw reads	Stand-alone version	Last update May 2020
Prophage Hunter (34)	3	Viral or phage genomes or contigs; host genomes for prophage prediction (FASTA files)	Web based (https://pro-hunter.genomics.cn/)	Last update Apr 2019
PhiSpy (40)	2	Viral or phage genomes or contigs	Stand-alone version	Last update May 2021
VIRALVERIFY (17)	3	Raw reads, phage or plasmid fragments from metagenomic assemblies	Stand-alone version	Last update 2020
ProphET (31)	1	Bacterial genome sequences	Stand-alone and web based (https://cpt.tamu.edu/galaxy-pub)	Self-updates or by the user
PHASTER (30)	1	Viral or phage genomes or contigs; host genomes for prophage prediction (FASTA files)	Web based (http://phaster.ca/)	Last update Dec 2020
Phigaro (32)	1	Metagenomic assemblies or raw genomes or contigs; host genomes for prophage prediction	Stand-alone version	Last update Aug 2020

[a]Corresponds to the approaches described in “Main Approaches.”

derived from phage genomes. Although those methods seem to be database independent, the ability of such prediction models to accurately identify phage sequences from a metagenomic assembly depends on the training data set, which is often based on a database such as the ones mentioned above. Furthermore, the training sequence sets must include genomic sequences from phages and bacteria, as well as other host-associated sequences such as plasmids (e.g., PPR-Meta, VirSorter2, and Prophage Hunter) (34–36). There are a variety of sequence features that different phage prediction methods focus on, including k-mer profiles, hits to protein sequence and/or HMMs databases, gene density, frequency of strand switch, and length of intergenic regions (36–39). For instance, PhiSpy employs customized AT and GC skew, k-mer information, strand orientation of transcripts, protein homology, and median protein length as inputs for its random forest classifier (40). Alternatively, some deep-learning methods generate prediction models based on different types of representations of the analyzed sequences. For instance, PPR-Meta uses “one-hot” encoding from the field of natural language processing to represent sequences as “base one-hot” and “codon one-hot” matrices (35).

Critical factors that affect the performance of phage and prophage prediction tools. As mentioned, the composition of the training data set is key to the performance of the phage prediction method. For instance, VirFinder’s prediction model was trained using a data set comprising bacterial, archaeal, and phage genomes from NCBI’s RefSeq database (37). Despite demonstrating great performance in the prediction of phage contigs from RefSeq genomes and simulated human gut metagenomes, the reported results revealed that the model’s performance varied depending on the domain (*Bacteria* or *Archaea*) or bacterial phylum targeted by the phages (37). Furthermore, a performance bias in VirFinder has also been reported in connection with the type of biome being surveyed, where it demonstrated a lower prediction performance for biomes that are poorly represented among the isolation sources of the phages in the training data set (41). Nonetheless, users have the option of retraining the prediction models using custom training data sets that include more phage and host sequences of interest (e.g., particular host taxa or specific biomes). For instance, the use of training sets that represented the bacterial and phage diversity in marine ecosystems significantly improved VirFinder's phage prediction performance (41). That being said, the second iteration of this tool, DeepVirFinder (42), has been shown to outperform VirFinder in identification of phage contigs when tested with viral RefSeq and human gut metagenomes.

In general, the way phage prediction tools perform may depend on whether the input data come from a total metagenome or a virus-enriched metagenome (virome). VirSorter, an example of predictors that follow approach 1, identifies phage sequences by comparing the values of a range of metrics between the sliding windows and a global value calculated for the complete set of input sequences (29). However, when the input data set corresponds to a virome, the calculated global values are not appropriate for discriminating between host and phage sequences. Thus, VirSorter allows the user to employ a set of precomputed global metrics to analyze input data sets enriched in viral sequences, by setting the –*virome* flag when using the command line version of the tool or by selecting the "virome decontamination" option if the tool is accessed via the CyVerse discovery environment (https://cyverse.org/) (29).

Another difference between the analyses of total metagenomes and viromes is the presence of contaminant cellular nucleic acids. Despite the viral purification process that precedes the generation of a virome, several studies have reported that the presence of contaminating bacterial sequences is rather common (43). Thus, such sequences should always be considered during the analysis of data sets derived from viral-enriched samples. Nevertheless, the abundance of contaminating sequences is generally higher in total metagenomes, and these must be correctly handled in order to minimize the number of unwanted sequences as input for the assembly, as these render the data analysis more complex and computationally expensive. Furthermore, the presence of contaminating sequences from eukaryotic organisms (host associated or members of the studied microbiome) can affect the performance of some phage prediction tools. For instance, it was recently reported that VirFinder's phage prediction model shows higher false-positive rates when the analyzed data sets contain sequences from eukaryotic organisms (43).

Selecting a Phage/Prophage Prediction Tool

One of the factors that the user should consider when choosing a tool is the level of computational expertise needed to perform the analysis. For users who are not experienced in the use of the command line, we recommend tools that are accessible via web browsers (e.g., PHASTER and Prophage Hunter). There are other tools that are available both as stand-alone versions for use on the command line and as web-based services hosted at genomics data analysis servers such as CPT Phage Galaxy (https://cpt.tamu.edu/galaxy-pub) and the CyVerse discovery environment (https://cyverse.org/). These include tools that are popular among the scientific community, such as VirSorter, VirSorter2, and VIBRANT, and the aforementioned data analysis servers offer user-friendly graphical interfaces that enable users who are not proficient at using the command line. However, for advanced users with more experience using the command line, we recommend the use of stand-alone tools, as these provide more flexibility by allowing the user to use custom reference databases and training data sets. Furthermore, advanced users may also combine different stand-alone tools into custom pipelines that produce a unified output based on the predictions reported by each tool. In fact, a publicly available pipeline named "What the Phage" can be downloaded from GitHub (https://github.com/replikation/What_the_Phage), and it allows users to apply a comprehensive set of phage prediction tools to their own metagenomic contigs and compare the output obtained from each one of them, which helps users to identify contigs that are reported by more than one tool and guide their selection of putative phage sequences (44). The use of stand-alone tools requires the user's capability to work on the command line in order to install and run these programs. Table 1 provides further details of some of the more commonly used tools for the identification of phages and prophages in metagenomic data sets.

Another important factor to consider during the selection of prediction tools is the extent of phage diversity that the user is interested in studying. For instance, an appropriate selection depends on whether the user is interested in analyzing only dsDNA phages or all phages regardless of their type of nucleic acid. Despite the massive increase in phage genomes that have been deposited in public databases during the last few years, there is a persistent bias toward genomes from dsDNA phages (45). In addition, most tools available to date use prediction models based on a specific set of features that represent all phage sequences, without considering the differences that could exist between different groups of phages. VirSorter2 stands out in this regard, as it is currently the only phage prediction tool that includes different prediction models for dsDNA, ssDNA, and RNA phages (36).

In addition to the type of nucleic acid, users may have an interest in particular groups of phages for which prediction tools have been developed. For instance, MG-Digger (46), Giant Virus Finder (47), and FastViromeExplorer (48) can identify nucleocytoplasmic large DNA virus (NCLDV) sequences in metagenomic data using nucleotide-level homology searches. However, it has been reported that average amino acid identity (AAI) between NCLDVs from different families can be as low as ∼20%. In consequence, ViralRecall (49) implements HMM searches from a database of 28,696 giant virus orthologous groups.

Assessing the completeness and contamination of predicted phage contigs. Following the prediction of phage contigs, the applied analysis pipeline might include a step that checks their quality in relation to genome completeness and host sequence contamination. VIBRANT, a phage predictor based on approach 3, includes a step in its algorithm that estimates genome completeness via the identification of terminal repeats (for circular genomes) and the presence of replication and viral hallmark proteins, detected though the search of the VOG database (http://vogdb.org/) (38). In addition, viralComplete determines the level of genome completeness by computing the similarity of the input phage contigs to each phage genome available at RefSeq and estimates whether the most similar RefSeq phage has a length similar to that of the input phage contig (17). However, the most popular and sophisticated tool developed to date for estimating genome completeness and host contamination of phage contigs is CheckV.

CheckV estimates host contamination in phage contigs through a method that combines gene annotation using a carefully curated HMM database of microbial and viral genes, and the detection of microbial or viral gene enriched regions using sliding windows (50). In addition, CheckV estimates the genome completeness of phage contigs through the identification of closely related phages in reference databases by means of the average amino acid identity (AAI) (50). Once a closely related phage is identified, the genome completeness of the input phage contig is estimated as the ratio of its length to that of the selected reference phage, and a confidence value is provided along with the genome completeness estimate (50). For cases in which the confidence value is low (i.e., when a sufficiently close reference phage is not identified for the input phage contig), CheckV estimates genome completeness using reference genomes that are annotated by the same set of viral HMMs as the input phage contig. Thus, in these cases CheckV provides a range of genome completeness values that correspond to the 5th and 95th percentiles from the distribution of reference genome lengths (50). Benchmarking of CheckV demonstrated enhanced performance in comparison with VIBRANT and viralComplete, and the output categories it employs are suitable for submitting phage genomic sequences to public databases, according to the Minimum Information about an Uncultivated Virus Genome standard (8, 50).

Contribution of phage and prophage prediction tools to the expansion of the known global phage diversity. The development of these prediction tools has greatly impacted the pace at which novel phage genome sequences are being discovered (45). For example, recent studies have explored the genomic characteristics of prophages (such as host range) in thousands of prokaryotic genomes (51). In these studies, the authors explored the prophage population using VirSorter alone or in combination with other tools, such as VIRALVERIFY. Moreover, the prophage prediction in large data sets is made in a rapid and easy way, allowing the analysis of other interesting aspects of prophages, for example, the identification of genes associated with virulence or antibiotic resistance (51).

In 2019, a study reported the viral exploration of the world's oceans through the analysis of a large collection of marine samples derived from the Global Ocean Viromes (GOV) 2.0 and Tara Oceans Polar Circle expeditions (52). The protocol for viral discovery applied in this study combined approaches 1 and 2 using VirSorter and VirFinder, respectively. This viral prediction protocol allowed the identification of 12 times more viral operational taxonomic units (OTUs) (≥10 kb in length) than the analysis conducted with data from the GOV 1.0 expedition. In addition, the applied protocol enhanced the identification of short viral contigs, <10 kb, which resulted in an additional 292,402 viral OTUs that had not been detected in GOV 1.0 (52). A similar phage prediction protocol was applied to contigs assembled from 28,060 gut metagenomic data sets, which resulted in the identification of 142,809 nonredundant phage sequences that were collectively referred to as the Gut Phage Database (GPD) (53). After clustering the GPD with phage sequences from other sources using a threshold of 90% nucleotide sequence identity over a 75% aligned fraction, the authors revealed that less than 1% of the resulting clusters contained entries from the GPD and NCBI's RefSeq database (53). Furthermore, after comparing GPD to the human Gut Virome Database (GVD) (54) and gut phages from IMG/VR (45), the authors demonstrated that GPD included the largest number of unique viral clusters, and thus, it significantly expanded the known diversity of human gut phages (53).

GENOME ANNOTATION OF PHAGES

Given the huge genomic diversity of phages, predicting the genes and functions encoded in their genomes is a key step that might provide better insight into their individual roles in their communities. While some viral genes are abundantly shared by a large proportion of the currently known viruses, many of their sequences have not been characterized yet. As an example, in the most recent iteration of the prokaryotic viral orthologous groups (pVOGs, formerly POGs), it was estimated that on average a third of the proteins in a dsDNA phage genome do not fall within any orthologous group in this database (33), suggesting that they have a novel function or belong to genes not commonly found in phages, likely being moved from the host. Furthermore, analyses of both human gut and environmental phage genomes indicate that around 75% of the sequences encoded by the phage genomes cannot be assigned to any biological function, suggesting that the functional novelty of phage sequences is even higher (7, 55). Additionally, phages tend to have a mosaic genome composition, with different genes having different evolutionary histories due to events of horizontal gene transfer (56), leading to further difficulties in the prediction process and selection of adequate reference databases to train the predictors. In order to characterize the functional diversity of phage genomes, two processes have to be carried out: (i) gene calling, or identifying the genes and their coordinates within the genome, and (ii) the functional annotation of those genes. Each of those processes entails different challenges and is worth exploring independently.

Gene Calling

The process of genome annotation begins with the identification of the genes present in a given phage. The most commonly and successfully applied tools to predict open reading frames (ORFs) in phages are Prodigal (57), Glimmer (58), and GeneMarkS (59, 60), even though they were initially developed for predictions in prokaryotes. Further improvement in gene calling accuracy has been observed when the above-mentioned tools are combined; thus, robust packages dedicated to a comprehensive genome annotation of microbial genomes, such as Rast-tk (61) (now part of PATRIC [62]) and Prokka (63), allow such tasks. Nevertheless, predicting ORFs in phages has its own challenges compared to prediction of ORFs in prokaryotes, and some genes might be missed by these tools and packages.

Overlapping genes, the presence of introns/inteins in phage genes, and alternative coding of phages add an extra layer to the complexity of gene calling. As the tools mentioned above do not automatically recognize these phenomena, the researcher must realize that the annotations may be incomplete or truncated. Here are some options that have recently emerged on how the community has approached each of these cases.

Overlapping genes. In phages, as in the entire viral spectrum, instances of gene overlap are prevalent (64, 65). Therefore, the nonmodularized genomic architecture of phages hinders gene prediction. To deal with that, McNair et al. developed PHANOTATE (66), a gene calling method specifically designed for phage genomes that outcompetes the other predictors in the number of predicted genes. Although some of the predictions might be false positives, given the high number of unknowns when working with phages, it might be preferable to initially have a relaxed tolerance for false positives.

Presence of inteins/introns. Many phages, such as T4 phages (67), members of the family *Herelleviridae* (68), and crAss-like phages (69), include group I or group II self-splicing introns/inteins. These selfish genetic elements in coding sequences cause fragmented gene calls and affect downstream steps such as functional annotation and phylogenetic analyses. Recently, Shapiro and Putoni developed Rephine, a pipeline for correcting gene calls (70). Briefly, given the pangenome of a related set of phages, Rephine evaluates two metrics, relbit (relative similarity of the gene fragments compared to the rest of the sequences in the cluster) and percoverlap (the overlap percentage between the gene fragments). Although the pipeline provides default values for these metrics, it also produces a table with the values obtained for each potential gene fragment. Therefore, it could be fruitful to explore and adjust the thresholds on a case-by-case basis, as the defaults are only suggested values based on the test cases examined by the authors.

Alternative genetic codes. Recently, the common use of alternative genetic codes in megaphages (71, 72) and crAss-like phages was reported (72). In these cases, regular gene calling results in fragmented ORFs and low coding density. To evaluate nucleic acid sequences for their genetic code, Dutilh and collaborators developed the tool FACIL (73). FACIL automates the detection of noncanonical codes; nevertheless, a manual inspection of the alignment of the codons on the DNA against homologous protein sequences is always a good idea. In cases where the use of an alternative genetic code is confirmed, the gene calling can be repeated using Prodigal, which supports all genetic codes defined by NCBI.

As for the vast majority of organisms, the researcher must be aware that automated gene calling on phages is not perfect; nevertheless, manual curation suffers from high labor cost, lack of standardization, and a degree of subjectivity in decision-making (74). Therefore, here are some final recommendations to reduce false positives and false negatives during gene calling. (i) All the above-mentioned "common ORF predictors" have a high degree of agreement. Use them in combination, selecting genes predicted by at least two tools (75, 76) and prioritizing Prodigal over Glimmer to assign start and end coordinates (74, 75). (ii) As mentioned, it is common to end up with truncated annotations and false negatives after the first round of gene calling. We encourage others to deal with the three possible causes addressed here, taking advantage of the mentioned tools to solve the problem in a semiautomated and case-based manner. (iii) Last, manually checking new findings reduces the number of false positives and allows the researcher to gain information on unique genomic features of the analyzed phages.

Annotation Approaches

After gene calling, different strategies for the functional annotation of ORFs can be applied depending on the genome or genomes in question. For common phages or phages with close homologues in public databases, query searches against sequence databases (Table 2) using BLAST (76) or DIAMOND (77) might be sufficient. Nevertheless, given the mosaicism and the high mutation rate of phages, often no significant results are obtained after a sequence similarity search. In this case, it is advisable to use methods for detection of remote homologs based on hidden Markov models (HMMs), which leverage the use of sequence profiles and the information about conservation for each

TABLE 2 Description of protein databases used for functional annotation of predicted ORFs

Database	Description	Type[a]
Viral RefSeq (149)	Curated NCBI database of viral genomes, genes, and proteins. Blast search is available online. Periodically updated.	Sequences
UniProtKB (150)	Curated collection of proteins and proteomes from all domains of life, derived from either direct submissions or predictions from either the European Nucleotide Archive (ENA), GenBank and the DNA Data bank of Japan (DDBJ). BLAST search is available online. Periodically updated.	Sequences
Pfam (151)	Database of protein families of all domains of life, derived from curated UniProtKB entries. Periodically updated.	MSA and HMM
viral eggNOG (79)	Clusters of orthologous viral proteins derived from graph-based unsupervised clustering. Last updated in 2016.	HMM
ViPhOGs (80)	Database of clusters of orthologous viral and phage protein domains generated through the CogSoft algorithm. Last updated in 2021.	HMM and MSA
pVOGs (33)	Phage gene families derived from orthologous clustering of phage proteins from complete phage genomes. Last updated in 2016.	HMM and MSA
NCBI_CD (152)	Collection of conserved bacterial domains, compiled from six different databases. Web search is available (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Last updated in 2020.	HMM and PSSM
SCOP (153, 154)	Database of structurally and evolutionary conserved proteins, organized in a hierarchical classification of families and superfamilies. Conserved domains based on different degrees of sequence identity are also available. Last updated in 2021.	Sequences
VOGdb (http://vogdb.org)	Database of clusters of orthologous viral proteins, derived from the combined use of the CogSoft algorithm and HH-suite on RefSeq phage and prophage genomes. The database provides both virus specific proteins (for detection of viral sequences in metagenomes), and panels of essential viral proteins. Periodically updated.	HMM and sequences
VPFs (155)	Database derived from the Earth Virome analysis of Paez-Espino et al. (7), consisting of groups of viral orthologous proteins. Last updated in 2016.	HMMs
PHROGs (81)	Remote homologous groups from proteins of complete genomes of viruses infecting bacteria or archaea. Last updated in 2021.	HMM and MSA

[a] MSA, multiple sequence alignment; PSSM, position specific scoring matrix.

residue. Programs from HMMER (http://hmmer.org/) or HH-suite (78) are commonly employed for this approach. Therefore, using them can be as easy as running BLAST either in the command line or in their online servers (HMMER, https://www.ebi.ac.uk/Tools/hmmer/; HH-suite, https://toolkit.tuebingen.mpg.de/tools/hhpred).

Multiple profile HMM databases can be used conjointly with HMMER or HH-suite (Table 2). Among them, we highlight pVOGs (33), viralOGs (now part of EggNOG) (79), ViPhOGs (80), and PHROGs (81), as they provide profiles of clusters of orthologous groups that are specific for phages and/or eukaryotic viruses. In all four cases, the raw protein alignments and the HMM profiles are provided so they can be used either with HMMER or HH-suite. Although not all profiles offered have a functional annotation, the identification of significant hits against those models implies identification of genes that are present and conserved in other viral genomes. Furthermore, these profile HMM databases can be used also to explore the evolutionary history of viruses and their proteins. For example, Low et al. employed pVOGs to build a concatenated protein phylogeny of dsDNA phages (82), and Andrade-Martínez et al. used the set of ViPhOGs used to elucidate potential relationships between *Herpesvirales* and *Caudovirales* and define the core genome of the former (83). Finally, although the use of profile HMMs is highly recommended for remote homolog searches, we recommend that users check their search results, since not all profiles have the same precision/sensitivity, and the same thresholds (E values) should not be applied for all profiles/genes.

Modular pipelines like Multiphate2 (84), RASTtk (61), and VIGA (85) allow coupling of gene calling and functional annotation in a high-throughput way using the command line. Furthermore, their modular construction approach permits the user to decide which parts to use. Multiphate2 can tackle previously mentioned phage-related annotation challenges, since it was designed to annotate and compare phage genomes. First, it integrates several gene callers, and second, it incorporates a variety of search algorithms and databases to increase the success rate of functional annotation.

Beyond the identification of genes and their functions, annotating a phage genome encompasses the annotation of RNAs, tandem repeats, and diversity-generating retroelements, among others. For example, RASTtk allows the annotation of RNAs and repeat regions. Although it was conceived to annotate bacterial and archaeal genomes, RASTtk has been a handy tool for phage annotation in the past (86). Moreover, while VIGA relies only on Prodigal for gene calling, its strength is that it includes several modules or routines to comprehensively annotate a viral genome. In addition to gene calling and functional annotation, VIGA allows detection of the contig shape (linear or circular) with LASTZ (87), prediction of ribosomal genes with INFERNAL (88), and prediction of tRNA and tmRNA sequences with ARAGORN (89) and runs PILER-CR (90), Tandem Repeats Finder (TRF) (91), and Inverted Repeats Finder (IRF) (92).

Machine learning-based annotation approaches. As mentioned before, most phage-borne genes have no close homologs in reference databases and their functional annotation is usually lacking. To solve this, several groups have used deep neural networks to recognize phage virion proteins, mainly because these can be considered a hallmark of phages and play an important role in phage-host interaction. DeepCapTail (93) classifies an ORF as capsid or tail; VirionFinder makes a binary classification of phage virion protein or not (94); DeephageTP focuses on Portal, TerL, and TerS proteins (95); and PhANNs classifies the proteins in 10 different classes of structural proteins (96). Ultimately, specific research and data requirements will lead to the selection of a specific classifier among those mentioned.

Selecting Gene Calling and Genome Annotation Tools

In general, we recommend the use of the mentioned modular pipelines, as they allow coupling of gene calling and functional annotation, the use of different programs, and the selection of the modules of interest. While all the pipelines require a basic knowledge of the command line, their installation is facilitated by the use of package managers; Multiphate2 relies on conda, VIGA relies on Docker, and RASTtk is now part of PATRIC (https://www.patricbrc.org/). In particular, for the overall selection of specific gene calling software, we suggest the use of a combination of approaches. The conjoined use of PRODIGAL, GLIMMER, and GeneMarkS provides a good overview of the main gene content of most phage genomes. Among these gene calling programs, only GeneMarkS offers a web server (http://opal.biology.gatech.edu/GeneMark/). Regarding the functional annotation, DIAMOND or BLAST can be used to search for homologs, and HMMER or HH-suite for remote homologs. All of them, except for DIAMOND, offer an online server to submit the sequences of interest.

Regarding the selection of a public database for genome annotation, we suggest the use of databases that are periodically updated and the application of both BLAST/DIAMOND and HMMs for annotation. This not only provides a higher confidence in the predicted functions of each sequence but also leverages the strengths of both approaches. In the case of BLAST/DIAMOND, the use of NCBI RefSeq is ideal in terms of both simplicity of running the process and number of viral sequences considered. The choice of an HMM database, in contrast, should be made based on the specific needs of the user, with PFAM being the most general option available, although the combination of multiple databases might yield the most useful results. In case no annotation is obtained via this combined approach, the use of machine learning-based tools might provide insights into the function of some of the unannotated phage ORFs.

TAXONOMIC CLASSIFICATION

The process of taxonomic classification of viral metagenomic samples is more challenging than that of cellular organisms. The shared common ancestry of the latter allows the existence of universal marker genes, most notably 16S and 18S rRNA genes, which can provide a reasonable representation of their evolutionary origin and divergence (1, 97). In contrast, this approach is not fully applicable to viruses, since they lack any equivalent set of universally conserved genes on which to construct a phylogeny (98, 99).

TABLE 3 Comparison of the tools for taxonomic classification of phage metagenomic data

Tool	Approach	Accessibility	Recently updated[a]
VIRIDIC (104)	BLAST followed by hierarchical clustering	Stand-alone version	Yes
VICTOR (105)	Blast-derived intergenomic distances	Online web service	No
VipTree (106)	tBLASTx	Online web service	Yes
Dougan and Quake method (107)	tBLASTx and 4-mer distances	Needs to be implemented by user	NA
VPF-Class (108)	HMMs against different databases	Stand-alone version	Yes
GRAViTy (110, 111)	Presence/absence and synteny of orthologous groups determined via HMMs	Stand-alone version	Yes
VIRify (https://GitHub.com/EBI-Metagenomics/emg-viral-pipeline)	HMMs from ViPhOGs	Online web service	Yes
Classiphage (97)	HMMs refined by BLASTp	HMM database available for download, distances must be self implemented	Yes
vConTACT (110)	Distances derived from Markov-based phage protein clusters	Stand-alone version	Yes

[a] "Yes" indicates that the tool or database has been updated or created in the past 2 years. NA, not applicable.

Viral classification is limited by two factors. First, current viral genome databases do not reflect the actual diversity of these elements in the biosphere (6). Second, viral taxonomy, as defined by the International Committee on Taxonomy of Viruses (ICTV), is currently in a state of flux (100), with changes being made and/or proposed to, among others, the number of taxonomic ranks to consider (101), the criteria for defining each specific rank (6), and the validity of traditional viral clades, not based on molecular criteria (102). However, several tools to classify viruses assembled from metagenomes have emerged in recent years (103), most of them based on sequence similarity by using either BLAST or HMMs. In this section, we briefly present some tools for classifying viruses from metagenomic data. Table 3 compares the main characteristics of these approaches. Other methods not mentioned here can be found in a recent comprehensive review by Nooij et al. (10).

BLAST-Based Approaches

Some of the tools developed for taxonomic assignment are based on the use of BLAST. For example, VIRIDIC (104) aligns all genomes (or contigs, for that matter) in a user-provided data set via BLASTn and then proceeds to utilize a hierarchical clustering algorithm to group the viral genomes based on their alignment similarities. In a similar manner, VICTOR, based on the application of Genome BLAST Distance Phylogeny (105), derives intergenomic distances from pairwise comparisons of nucleotide or amino acid sequences from complete or partial viral genomes. It is important to note that both VICTOR and VIRIDIC find similarities between the genomes or contigs which are introduced as input, either clustering them or constructing a phylogeny, but do not compare them to any database or set of genomes with known taxonomy. In other words, they do not directly classify phage genomes into specific clades. In order to use them for classification, the user must add reference viral genomes, with known taxonomy, to the input and use their similarity to the original input genomes to determine their classification. For VIRIDIC, the working similarity thresholds available in the tool only allow the generation of putative species or genus clusters, according to recent definitions by ICTV. Another option is to perform an extra analysis to determine the clades associated with each cluster of genomes via identification of shared gene orthologous groups representative of specific viral clades. Both tools are restricted in the maximum number of sequences to use: 100 for VICTOR and 300 for VIRIDIC.

tBLASTx is used by VipTree (106), a software tool designed for quick taxonomic classification of new viral genomes. VipTree determines pairwise similarities of putative proteomic sequences of a query genome against reference viral genomes included in the tool and generates a phylogenetic tree based on said similarities. The software is limited to constructing trees with at most 200 query sequences.

The usefulness of these three tools is limited by the degree of completeness of the query viral contigs, the prior knowledge the author might have regarding their taxonomy, and the database employed. VICTOR, given its use of nucleotide distances, is better suited for determining close evolutionary relationships, while VIRIDIC and VipTree, given their ability to use proteomic information, are better for more divergent ones. In the extreme case where little or nothing is known of the taxonomy of the queried viruses, the use of reference sequences automatically provided by VipTree would be preferred to the manual input of reference sequences from VICTOR or VIRIDIC.

In fact, tBLASTx has been shown to be able generate large-scale viral clusterings when combined with k-mer information: Dougan and Quake (107) employed a combination of tBLASTx and 4-mers to derive an entropy-based genomic distance for classifying a set of 5,817 viral genomes. These distances are then converted into a multidimensional representation through t-distributed stochastic neighbor embedding (t-SNE) and clustered to produce a distance dendrogram. Albeit done with complete viral genomes, this method could be adapted for metagenomic data sets by an experienced user and is known to work with a large set of viral genomes, compared to the three aforementioned tools.

Markov-Based Approaches

Other approaches use HMMs and/or Markov clustering (MCL) for classification (either conjointly with or without BLAST). For example, VPF-Class (108) runs an HMM search against three different data sets: the NCBI viral sequences, the prophages data set from Roux et al. (29), and the Global Ocean Virome (109) to determine viral classification and host prediction.

In contrast, GRAViTy (110, 111) derives protein profile HMMs from BLASTp-based orthologous protein clusters. These models are then used to scan complete viral genomes to derive information of their gene content and orientation (i.e., synteny), from which pairwise distances are computed. GRAViTy has been applied for classification of both complete eukaryotic (110) and prokaryotic (111) viruses. As mentioned in the section above, the use of presence/absence information of ViPhOGs (formerly VDOGs) in a given set of viral genomes, obtained via scanning of genomic proteins against ViPhOG HMMs, can be used both to infer phylogenies among viruses and to determine representative orthologous proteins of specific viral taxa, allowing the potential classification of viral sequences from metagenomic samples (80, 83). VIRify (https://github.com/EBI-Metagenomics/emg-viral-pipeline), currently in development, employs ViPhOG presence/absence information for the classification of metagenomic viral contigs.

Another tool, Classiphage, leverages a combined approach: HMMs are constructed from the proteomes of a set of genomes, clustered based on Markov chain clustering. BLASTp is then employed to refine the clusters, with the objective of finding unique HMMs for each cluster of interest (97). However, this tool has been tested in only a very specific set of phage genomes, namely, vibriophages (i.e., phages which infect bacteria from the genus *Vibrio*).

Finally, vConTACT, currently in its second version (110), constructs phage protein clusters via Markov clustering and generates pairwise similarities between genomes. These similarities can then be represented as a gene-sharing network from which genome clusters can be derived. However, the user needs to use reference viruses to assign the taxonomy.

Either VIRify or VPF-Class would be preferably used for classification of phage contigs when no taxonomic background is known. GRAViTy, while leveraging both HMMs and synteny, can achieve reliable classification only up the family level (albeit with high accuracy), and its use is not recommended for incomplete (or putatively incomplete) genomes. vConTACT, on the other hand, would be advisable for looking into evolutionary relations between large sets of phages with different degrees of evolutionary distances, where some information is known about the taxonomy of the genomes, such that adequate reference sequences can be included in the analysis.

Selecting a Taxonomic Classification Tool

A critical aspect to consider when selecting a tool to use is the computational expertise and software support. VICTOR, GRAViTY, and VipTree are available as online web services, such that the user needs no command line experience to run them. On the other hand, GRAViTy, VIRIDIC, VPF-Class, VIRify, and vConTACT offer stand-alone versions which can be included in custom pipelines to analyze large data sets derived from metagenomic data.

The installation and running of the programs vary in accessibility: VIRIDIC is very straightforward to install and run. GRAViTy, VPF-Class, and vConTACT are based on Python and therefore might require basic knowledge of both Python programming language and command line usage. Notably, vConTACT provides multiple installation options, all of them explained by the authors, although it requires the installation of dependencies prior to its usage. VPF-Class needs to be compiled by the user before running it. VIRify is installed via Nextflow and either Docker or Singularity.

All of these software programs provide documentation for download, installation, and use, and most are available in GitHub, where some support may be obtained from the developers. Of these approaches, VPF-Class, VIRify, VIRIDIC, GRAViTy, and vConTACT have been updated (or created, for that matter) in the past 2 years.

The procedure described by Dougan and Quake (107) is not available in any sort of software or source, so the user must replicate their workflow as described in the paper. A similar process must be followed for Classiphage, where the user must download the associated vibrophage HMMs and follow the methods described by the authors.

Overall, we suggest the combination of one Markov-based and one BLAST based procedure, to have complementary data for classification. For users without computational experience, GRAViTY and VipTree, both with web services, would be the easiest combination to run. Users with a basic knowledge of the command line interface might prefer a combination of VipTree and VPF-Class, or even VipTree and VIRify. For highly experienced users, especially with large data sets, the prior analyses could be complemented with the use of vConTACT, to obtain a large-scale picture of the taxonomic composition and relations of the phages in the metagenome.

PHAGE-HOST INTERACTIONS

Since next-generation sequencing (NGS) technologies are culture independent, viral sequences identified from these methods lack an association with their host (111). While many experimental methods have been developed to link a phage with its bacterial host, these methods are not always applicable, and they are often biased toward culturable hosts, as they require the purification of the host, phage, or both (111). Consequently, there is an urgent need for computational tools which predict phage-host interactions.

Tools for detecting phage-host interactions are helping microbial ecologists seeking to solve questions related to phage biology and their interaction with other members of microbial communities. Additionally, they give clues to the potential host of a newly discovered virus, or the potential viruses to which an unidentified bacterium is susceptible (111–113). Moreover, these tools may be a source of new knowledge, as they can predict potential interactions that have not been described before. However, it is important to be careful and select the analysis that best suits specific research questions in regard to the balance of false-positive and false-negative interactions and which is prioritized.

Main Approaches

A recent increase in sequence-based tools aimed at identifying which bacteria act as the hosts of a given phage in a metagenomic sample has been observed (112, 114–119). Some of these tools use prediction signals from phage-host interactions that can be categorized as (i) homology dependent, such as nucleotide similarity BLAST scores and CRISPR spacer matches (111), as in SpacePHARER (119), CrisprOpenDB (120) and

CRISPRDetect+BLAST (121), or (ii) independent signals, such as similarity/dissimilarity measures, oligonucleotide (k-mer) usage profiles, and protein-protein and domain-domain interaction scores, as in VirHostMatcher (112), RaFAH (113), WIsH (114) and HostPhinder (115). Other tools, such as Host Taxon Predictor (HTP) (116), ILMF-VH (117) and VirHostMatcher-Net (118), also include features obtained from the genomic sequence itself, like the length and molecule type (ssDNA, dsDNA, ssRNA, etc.), or even features derived from virus-virus and host-host similarity, assuming that similar phages may infect the same host, and that genetically similar hosts are susceptible to the same type of phage (Table 4).

In a similar manner to homology-search based tools looking for prophages in bacterial genomes, phage host predictions generated by CRISPR-based homology-dependent tools, such as SpacePHARER, the CrisprOpenDB pipeline, and CRISPRDetect+BLAST, tend to be highly specific, producing a low rate of false positives, but not very sensitive, generating a high number of false negatives (Table 4). The presence of CRISPR spacers reflects past phage infections, which are signals of phage-host interactions. However, CRISPR spacer content within bacterial genomes is dynamic, meaning the same set of spacers is not necessarily conserved over time or among different strains (122). Moreover, not all bacteria use CRISPR spacers to overcome phage infection: as of 2021, only 63.3% of the bacterial genomes available at NCBI databases have CRISPR spacers (120). Therefore, considering database biases and CRISPR spacer dynamics by themselves, there are still a number of interactions that cannot be identified by this method.

In general, homology-independent tools use as their input full genomic sequences or assembled contigs from the phage, the bacterial host, or both, to feed a previously trained machine learning model which predicts phage-host associations in a taxonomically independent manner (112, 114–117, 123). Conversely, tools such as vHULK (124) and HostPhinder, which use their own databases to make predictions, might bias the results toward the taxa included in said databases. Altogether, these software tools have moderate to high predictive ability for all taxonomic levels, from domain to species, with the lowest performances being obtained for the lowest taxonomic levels (i.e., genus and species).

Table 4 summarizes the main tools for prediction of phage-host interactions and shows the performance obtained for each tool, although caution should be taken in interpreting those numbers, as they are derived from different benchmark data sets and for different taxonomic levels. It is important to highlight that some tools do not report performance metrics other than accuracy, which skews the perception of how well the model works. Usually, phage-host interaction data sets are imbalanced toward negative interactions (i.e., when there is no evidence that a phage is able to infect a given bacterium), meaning that there is a higher number of negative interactions than positive ones. Thus, if the model predicts that every interaction is negative, the accuracy will display a high value, causing the user to think the model is good. Therefore, it is better to rely on other metrics, such as F1 score and the area under the receiver operating characteristic curve (AUC) of precision recall and/or sensitivity versus specificity curves (125), which do not consider true negatives as part of their calculation.

Tools such as RaFAH, ILMF-VH, PHISDetector (123), and HTP, based on machine learning techniques such as random forest, logistic regression, support vector machines, etc., show better performance than those that are directly metric based, for example, those that predict the host as the one with the highest similarity or lowest dissimilarity value, such as VirHostMatcher and HostPhinder, or those that are homology dependent, such as those based on BLAST or CRISPR matches (119–121). Additionally, vHULK and VirHostMatcher-Net, which include network-based models, such as k-nearest neighbors, may be negatively affected when the query is distant from known viruses.

Selecting a Host Prediction Tool

Most of these tools (i.e., SpacePHARER, CrisprOpenDB, WIsH, Host Taxon Predictor, ILMF-VH, and VirHostMatcher-Net) are available from GitHub repositories and have

TABLE 4 Model performance metrics per phage-host prediction tool obtained using their own benchmark data sets, and description of each prediction method[a]

Phage-host prediction tool	Prediction method[b]	Taxonomic level	AUC	F1 score	Accuracy
SpacePHARER (119)	Identifies CRISPR spacers in bacterial genomes, translates those into protein motifs, and performs a protein alignment against possible protospacer motifs from phage sequences; the prediction is based on a probability score which selects the host with the higher likelihood.	Phylum			0.87
		Class			0.84
		Order			0.8
		Family			0.79
		Genus/species			0.77
CrisprOpenDB pipeline (120)	Looks for CRISPR spacer matches (alignments) and applies some filters (no. of gaps, position of the spacer in the bacterial genome, and taxonomic accordance between predictions) to make predictions at the lowest taxonomic levels possible.	Genus		0.57	
WIsH (114)	Computes the maximum likelihood of phage P infecting host H based on the training of a homogeneous Markov model.	Genus			0.35
		Family			0.43[c]
		Order			0.48[c]
		Class			0.6[c]
		Phylum			0.75[c]
VirHostMatcher (112)	Calculates different dissimilarity/similarity measures based on genomic composition to predict the host of a given phage. The least dissimilar or the most similar measure indicates a likely positive interaction.	Genus			0.33
		Family			0.48
		Order			0.54
		Class			0.67
		Phylum			0.75
HostPhinder (115)	Predicts the host of a phage as the host of the most genomically similar phage present in the reference database. Genomic similarity refers to how much of the reference is covered by the query sequence.	Species			0.74
		Genus			0.81
Viral Host UnveiLing kit (vHULK) (124)	Uses protein annotation alignment scores to pVOGs database to construct the matrix to predict interaction using a neural network.	Species			0.52
		Genus			0.82
Random Forest Assignment of Hosts (RaFAH) (113)	Uses the scores obtained from the search of protein clusters in viral genomes to predict interaction using a random forest model.	Genus		0.67	
		Family		0.75	
		Order		0.78	
		Class		0.81	
		Phylum		0.91	
		Domain		0.99	
ILMF-VH (117)	Kernelized logistic matrix factorization applied on a virus-virus network, a host-host network, or a virus-host network. Each network is computed based on oligonucleotide frequencies, Gaussian interaction profile, and previously known virus-host interactions.	Species	0.92		0.64

(Continued on next page)

TABLE 4 (Continued)

Phage-host prediction tool	Prediction method[b]	Taxonomic level	AUC	F1 score	Accuracy
VirHostMatcher-Net (118)	Uses homology-dependent and -independent features to compare between virus and hosts: CRISPR matches, homology scores from BLAST, an alignment-free similarity measure, WIsH maximum likelihood, comparison between viruses that infect the same host (SV+) and viruses infecting different hosts (SV−). The prediction is based on a Markov random field model.	Species			0.43
		Genus			0.59
		Family			0.7
		Order			0.78
		Class			0.83
		Phylum			0.86
PHISDetector (123)	Uses 6-mer frequencies, codon usage, prophage detection, CRISPR matches and protein-protein interactions information to predict interactions. The model trained for the prediction is an ensemble of different machine learning models (random forest, decision tree, logistic regression, SVMs with different kernel settings, Gaussian and Bernoulli naive Bayes).	Genus or species	0.93	0.88	0.88
Host Taxon Predictor (HTP) (116)	Uses nucleotide sequence information such as molecule type, mono/dinucleotide absolute frequencies and di/trinucleotide relative frequencies to predict if a virus is phage or nonphage. They have implemented 4 types of classifiers: logistic regression, quadratic discriminant analysis, support vector machines, and k-nearest neighbors.	Domain	0.98		0.93

[a] Note that the performance metrics are not directly comparable, as different tools were benchmarked with data sets which might or might not be the same.
[b] SVMs, support vector machines.
[c] Contig-based prediction.

documentation available for installation and implementation. All these tools, with the exception of ILMF-VH (last updated in 2019), have been updated over the course of the past year (2020).

RaFAH and HostPhinder are available on SourceForge and Docker, respectively. PHISDetector and HostPhinder also have web tools available, so these tools may be easier to apply than the others for users with no previous experience with command lines. As mentioned above, we recommend using tools trained on supervised machine learning methods such as RaFAH and PHISDetector when dealing with data derived from complex communities and complementing the analysis with CRISPR similarity-based tools (120), as these will allow the researcher to gather more information from the data.

Phage Lifestyle Prediction

In addition to determining whether a phage can interact with a given bacterium, in microbial communities it is also relevant to identify the type of interaction they have. Phages can follow many types of different life cycles: (i) lytic, in which they lyse their cells after using the host machinery to replicate; (ii) temperate, in which phages integrate into the bacterial genome or remain as a circularized DNA portion outside the bacterial chromosome, propagating each time the bacterial host replicates; (iii) chronic, similar to the lytic cycle but with the distinction that the phage is not able to lyse the cell; and (iv) pseudolysogenic, in which the phage remains dormant without propagating or integrating within the bacterial genome (126). With the lytic and temperate cycles being the most extensively characterized cycles in nature, there exist random forest-based tools which use protein similarity scores, such as PHACTS (127), or conserved protein domain presence, like BACPHILIP (128), to predict the lifestyle of a phage sequence. While these tools have shown good performance, with an F1 score of >0.8 (128), it is important to note that they were trained on limited data sets, biased to culturable phage-host pairs, whose interactions are very well described. Therefore, these tools may not be very accurate for use with novel or distant phage sequences.

MICRODIVERSITY

Microbial populations experience constant genetic changes. Besides being important drivers for bacterial evolution, given their capability to promote horizontal gene transfer, phages undergo several changes within their genomes, leading to a wide collection of mutations that are chosen by natural selection (129). Moreover, the high substitution rates shown by some viruses in a short time period increase the number of significant changes within phage genomes, allowing speciation evens and, therefore, the diversification of the ecosystem (130). It is known that the intrapopulation genetic variation, or microdiversity, of viruses has an impact in shaping host dynamics. Furthermore, changes related to an increase in the infection rates and host range might produce bacterial changes related to resistance against viral infection, generating antagonistic evolution dynamics (131).

Antagonistic changes have been reported in multiple environments, such as the human gut (130–132) and marine ecosystems (133). These have been strongly associated with the maintenance of ecological microbial equilibrium within the human gut but also with its diversification. Given that most alterations within phage genomes are related to adaptations to new host strains or the ability to infect strains that are resistant to the infection, an increase in phage populations with new genomic features leads to the infection of bacterial strains that are usually found in high abundance. This infection produces a decrease in bacterial population but also a selection pressure that allows the host to acquire resistance. This pattern leads to the regulation of bacterial populations and an increase in bacterial adaptation to their new predators (134). Analogous to this process, elements such as the cyanophages undergo mutations depending on how optimal their host is. When infecting optimal hosts, phages accumulate few mutations within their genomes. In the opposite case, in which the host is suboptimal, the number of mutations increases and so does the diversity of viral populations (135).

Independently of the driver, genomic mutations in viruses from the human gut have been found to be common in healthy individuals. As described by Minot et al. (130), even though 80% of viral elements are shown to remain stable over long periods of time, mutations start to rapidly accumulate within their genomes, driving interindividual variation. In marine environments, the frequencies of polymorphic sites and nonsynonymous mutations within housekeeping genes have been found to be depth dependent, highlighting the importance of genomic variations on modulating viral functions (136).

Main Approaches

Despite the importance of the dynamics driven by intraspecific variation on different environments, microdiversity analysis is considered a recent approach in virome analysis. Therefore, the development of specific tools to characterize viral microdiversity is still an ongoing process, with few tools already available. Consequently, studies regarding single nucleotide polymorphism (SNP) calling, calculation of base substitutions, and virome stability within individuals have had to rely on general software and frameworks for the discovery of variants from NGS data. As proposed by DePristo et al. (137), a framework for variant calling and genotyping must include at least 3 phases: (i) manipulation of NGS data, consisting of mapping the reads against the genomes or contigs of interest, followed by a series of alignment refinement and base quality recalibration; (ii) a variant discovery step, in which SNPs, structural variations (SV), and indels are reported, followed by a genotyping process; and (iii) association analysis and computation of microdiversity metrics. When applied to virome data, the variant calling process must rely on a well-curated set of contigs, which are expected to have a minimum sequencing depth of 5× (134, 138) or 10× (130, 133, 139). Furthermore, in order for a SNP to be considered valid, it must be supported by at least 4 reads (140). Likewise, a valid SNP will also depend on the quality call threshold, defined as a Phred-based value representing the confidence of the presence of a specific variation in a given site (141). However, to the best of our knowledge, there is no universal reference value for this metric, implying that the threshold is chosen based on specific characteristics for each analysis.

To be able to make biological inferences with sequence variant data, multiple metrics can be applied. To measure the degree of polymorphism in a given population, the nucleotide diversity (π) metric, understood as the average number of differences found within a specific region of DNA from two different taxa, can be implemented. As explained by Schloissnig et al. (140), for estimating π between a pair of metagenomic samples, a variation of the original formula proposed by Begun et al. (142) is made, in order to include the possibility of having more than two alleles per site. In this case, the chance of choosing different alleles in a specific and randomly chosen position in the genome is computed. On the other hand, the fixation index (Fst) is also used for measuring population differentiation with no changes to the original formula. For analyzing the effect of natural selection in the virome population, a measure of the ratio between nonsynonymous and synonymous substitutions is calculated (*pN*/*pS*), if all mutations have the same probability of occurring across the genome (52).

Even though few specific pipelines for viral microdiversity are available at the moment, those available allow performing the multiple steps involved in a normal analysis in a straightforward and user-friendly way. For example, MetaPop is described as a tool for the manipulation and visualization of micro- and macrodiversity-related data. In fact, it can be run completely via a single command. Furthermore, it implements multiple metrics, including the ones described above, but also Watterson's theta nucleotide diversity, Tajima's D, and codon usage biases (143). Among the pipelines designed exclusively for viral sequences, DiversiTools (http://josephhughes.github.io/DiversiTools/), which was initially developed for the analysis of eukaryotic viruses, can also be used for virome data in general, by applying p*N*/p*S* metrics for comparisons between samples (136). In contrast to the previous tools, the inStrain pipeline is considered a suite of programs for analyzing metagenomic data which can be easily applied to infer exclusively viral dynamics (138) with metrics such as nucleotide diversity, linkage, p*N*/p*S*, and iRep (measure of how fast a population is replicating) and a user-friendly environment (144).

Microdiversity within viral populations is an important factor when assessing the dynamics related to the virome of specific ecosystems. These analyses are extremely useful in longitudinal studies when mutation accumulation rates and temporal succession are used to evaluate long-term stability within a population (130, 132, 133). Even though software and frameworks have been adapted for viral communities, new tools have been developed for making this process faster and more reproducible, which might lead to an increase in the number of studies regarding this topic in the future.

Selecting a Microdiversity Tool

The decision regarding how to design a microdiversity analysis should prioritize the bioinformatic expertise of the researcher, rather than the performance of the available tools, since most of the reported workflows are based on alignment and SNP-calling software tools that have been benchmarked several times, yielding adequate results. For generating microdiversity statistics, microbial diversity packages such as Vegan (145) and Phyloseq (146) can be applied for completing the workflow, but the multiple changes or differences in the in-house scripts developed for implementing those packages might lead to reproducibility problems. The standardized pipelines described above can reduce the chance of making the analysis unreproducible and, therefore, can be implemented for running the entire analysis or for comparing the results obtained by the different in-house scripts. Regarding automatic pipelines (i.e., MetaPop and inStrain), while we consider that both perform well for generating microdiversity inferences and also provide a user-friendly environment for installing and running the tools, the number of metrics available on MetaPop makes it slightly more appealing for microdiversity analyses.

CONCLUDING REMARKS

We have presented a brief overview of the tools used for the principal analyses of phage metagenomic data, namely, assembly and detection of phage and prophage sequences, annotation, taxonomic classification, host identification, and microdiversity. As we have shown, the choice of a specific tool must consider not only the approach employed but also accessibility and the data the researcher is working with, as well as the degree of support from the developers of the software. As more software is integrated into metagenomic pipelines, the use of these tools should become easier for the user, regardless of their computational expertise. Moreover, such integration will allow the simultaneous use of multiple tools, employing different approaches, to address specific metagenomic problems (e.g., taxonomic classification) in a given phage data set, allowing the cross-examination of their outputs to determine which findings are sustained between different tools.

Of notable importance is the growth over recent years in the total number of machine learning-based tools for different virome procedures. Given the reliance of these tools in their training data sets, the constant increase in virome data, both simulated and real, should provide us with better benchmarks of their performance and therefore lead to an overall improvement of their ability to characterize metagenomic data. Parallel improvements might also be expected from other database-reliant but non-machine learning tools, such as BLAST- or HMM-based annotation or classification software.

It is also relevant to acknowledge the scarcity of tools and knowledge regarding ssDNA and RNA phages. Despite the fact that both types of phages have been shown to be highly abundant in a variety of environments, limitations in experimental protocols for viral isolation and lack of sufficient ssDNA and RNA genomes in databases have led to a reduced number of isolated sequences (19, 147, 148). The recent work of Callanan et al. in identification and assembly of ssRNA viruses (148), as well as improvements in experimental techniques and sequencing technologies, provides an exciting outlook for tapping the diversity of these less known phage clades.

ACKNOWLEDGMENTS

We thank Mariana Piuri for her aid in revising the manuscript.

G.L.-L. received a postdoctoral fellowship (2019-000012-01EXTV-00488) from CONACyT. L.F.-J. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 801199.

REFERENCES

1. Garza DR, Dutilh BE. 2015. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell Mol Life Sci 72:4287–4308. https://doi.org/10.1007/s00018-015-2004-1.
2. Fancello L, Raoult D, Desnues C. 2012. Computational tools for viral metagenomics and their application in clinical research. Virology 434: 162–174. https://doi.org/10.1016/j.virol.2012.09.025.
3. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15:161–168. https://doi.org/10.1038/nrmicro.2016.177.
4. Edwards RA, Rohwer F. 2005. Viral metagenomics. Nat Rev Microbiol 3: 504–510. https://doi.org/10.1038/nrmicro1163.
5. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, Stekel DJ, Hobman J, Jones MA, Millard A. 2021. INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. Phage (New Rochelle) 2:214–223. https://doi.org/10.1089/phage.2021.0007.
6. Dion MB, Oechslin F, Moineau S. 2020. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 18:125–138. https://doi.org/10.1038/s41579-019-0311-5.
7. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. Nature 536:425–430. https://doi.org/10.1038/nature19094.
8. Cantalupo PG, Pipas JM. 2019. Detecting viral sequences in NGS data. Curr Opin Virol 39:41–48. https://doi.org/10.1016/j.coviro.2019.07.010.
9. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee K-B, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit M-A, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, et al. 2019. Minimum information about an uncultivated virus genome (MIUViG). Nat Biotechnol 37:29–37. https://doi.org/10.1038/nbt.4306.

10. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MP. 2018. Overview of virus metagenomic classification methods and their biological applications. Front Microbiol 9:749. https://doi.org/10.3389/fmicb.2018.00749.
11. Zheng T, Li J, Ni Y, Kang K, Misiakou MA, Imamovic L, Chow BK, Rode AA, Bytzer P, Sommer M, Panagiotou G. 2019. Mining, analyzing, and integrating viral signals from metagenomic data. Microbiome 7:42. https://doi.org/10.1186/s40168-019-0657-y.
12. Labonté JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J 7:2169–2177. https://doi.org/10.1038/ismej.2013.110.
13. Sutton TD, Clooney AG, Ryan FJ, Ross RP, Hill C. 2019. Choice of assembly software has a critical impact on virome characterisation. Microbiome 7:12. https://doi.org/10.1186/s40168-019-0626-5.
14. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116.
15. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
16. García-López R, Vázquez-Castellanos JF, Moya A. 2015. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. Front Bioeng Biotechnol 3:141. https://doi.org/10.3389/fbioe.2015.00141.
17. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. Metaviral SPAdes: assembly of viruses from metagenomic data. Bioinformatics 36:4126–4129. https://doi.org/10.1093/bioinformatics/btaa490.
18. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8:giz100. https://doi.org/10.1093/gigascience/giz100.
19. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. 2020. Expansion of known ssRNA phage genomes: from tens to over a thousand. Sci Adv 6:eaay5981. https://doi.org/10.1126/sciadv.aay5981.
20. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B. 2019. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ 7:e6800. https://doi.org/10.7717/peerj.6800.
21. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TP, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17:1103–1110. https://doi.org/10.1038/s41592-020-00971-x.
22. Zablocki O, Michelsen M, Burris M, Solonenko N, Warwick-Dugdale J, Ghosh R, Pett-Ridge J, Sullivan MB, Temperton B. 2021. VirION2: a short-and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. PeerJ 9:e11088. https://doi.org/10.7717/peerj.11088.
23. François S, Filloux D, Frayssinet M, Roumagnac P, Martin DP, Ogliastro M, Froissart R. 2018. Increase in taxonomic assignment efficiency of viral reads in metagenomic studies. Virus Res 244:230–234. https://doi.org/10.1016/j.virusres.2017.11.011.
24. Arisdakessian CG, Nigro OD, Steward GF, Poisson G, Belcaid M. 2021. CoCoNet: an efficient deep learning tool for viral metagenome binning. Bioinformatics 37:2803–2810. https://doi.org/10.1093/bioinformatics/btab213.
25. Johansen J, Plichta D, Nissen JN, Jespersen ML, Shah SA, Deng L, Stokholm J, Bisgaard H, Nielsen DS, Sørensen S, Rasmussen S. 2021. Genome binning of viral entities from bulk metagenomics data. bioRxiv https://doi.org/10.1101/2021.07.07.451412.
26. Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJ, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. 2021. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 39:555–560. https://doi.org/10.1038/s41587-020-00777-4.
27. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165. https://doi.org/10.7717/peerj.1165.
28. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:7:e7359. https://doi.org/10.7717/peerj.7359.
29. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985. https://doi.org/10.7717/peerj.985.
30. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21. https://doi.org/10.1093/nar/gkw387.
31. Reis-Cunha JL, Bartholomeu DC, Manson AL, Earl AM, Cerqueira GC. 2019. ProphET, prophage estimation tool: a stand-alone prophage sequence prediction tool with self-updating reference database. PLoS One 14:e0223364. https://doi.org/10.1371/journal.pone.0223364.
32. Starikova EV, Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Ilina EN, Govorun VM. 2020. Phigaro: high-throughput prophage sequence annotation. Bioinformatics 36:3882–3884. https://doi.org/10.1093/bioinformatics/btaa250.
33. Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res 45:D491–D498. https://doi.org/10.1093/nar/gkw975.
34. Song W, Sun H-X, Zhang C, Cheng L, Peng Y, Deng Z, Wang D, Wang Y, Hu M, Liu W, Yang H, Shen Y, Li J, You L, Xiao M. 2019. Prophage Hunter: an integrative hunting tool for active prophages. Nucleic Acids Res 47:W74–W80. https://doi.org/10.1093/nar/gkz380.
35. Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, Zhu H. 2019. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. GigaScience 8:giz066. https://doi.org/10.1093/gigascience/giz066.
36. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB, Roux S. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9:37. https://doi.org/10.1186/s40168-020-00990-y.
37. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5:69. https://doi.org/10.1186/s40168-017-0283-5.
38. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90. https://doi.org/10.1186/s40168-020-00867-0.
39. Amgarten D, Braga LP, da Silva AM, Setubal JC. 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. Front Genet 9:304. https://doi.org/10.3389/fgene.2018.00304.
40. Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res 40:e126. https://doi.org/10.1093/nar/gks406.
41. Ponsero AJ, Hurwitz BL. 2019. The promises and pitfalls of machine learning for detecting viruses in aquatic metagenomes. Front Microbiol 10:806. https://doi.org/10.3389/fmicb.2019.00806.
42. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. 2020. Identifying viruses from metagenomic data using deep learning. Quant Biol 8:64–64. https://doi.org/10.1007/s40484-019-0187-4.
43. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol 3:130160. https://doi.org/10.1098/rsob.130160.
44. Marquet M, Hölzer M, Pletz MW, Viehweger A, Makarewicz O, Ehricht R, Brandt C. 2020. What the Phage: a scalable workflow for the identification and analysis of phage sequences. bioRxiv. https://doi.org/10.1101/2020.07.24.219899.
45. Roux S, Páez-Espino D, Chen IM, Palaniappan K, Ratner A, Chu K, Reddy TB, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides N. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 49:D764–D765. https://doi.org/10.1093/nar/gkaa946.
46. Verneau J, Levasseur A, Raoult D, La Scola B, Colson P. 2016. MG-Digger: an automated pipeline to search for giant virus-related sequences in metagenomes. Front Microbiol 7:428. https://doi.org/10.3389/fmicb.2016.00428.
47. Kerepesi C, Grolmusz V. 2017. The "Giant Virus Finder" discovers an abundance of giant viruses in the Antarctic dry valleys. Arch Virol 162:1671–1676. https://doi.org/10.1007/s00705-017-3286-4.
48. Tithi SS, Aylward FO, Jensen RV, Zhang L. 2018. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. PeerJ 6:e4227. https://doi.org/10.7717/peerj.4227.
49. Aylward FO, Moniruzzaman M. 2021. ViralRecall—a flexible command-line tool for the detection of giant virus signatures in 'omic data. Viruses 13:150. https://doi.org/10.3390/v13020150.
50. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–585. https://doi.org/10.1038/s41587-020-00774-7.

51. López-Leal G, Camelo-Valera LC, Hurtado-Ramírez JM, Verleyen J, Castillo-Ramírez S, Reyes-Muñoz A. 2021. Mining of thousands of prokaryotic genomes reveals high abundance of prophage signals. bioRxiv https://doi.org/10.1101/2021.10.20.465230.
52. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay J, Vik D, Coordinators T, Acinas SG, Babin M, Bork P, Boss E, Bowler C, Cochrane G, Vargas C, Follows M, Gorsky G, Grimsley N, Guidi L, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Karp-Boss L, Karsenti E, Not F, Ogata H, Pesant S, Poulton N, Raes J, Sardet C, Speich S, Stemmann L, Sullivan MB, Sunagawa S, Wincker P, Babin M, Tara Oceans Coordinators, et al. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177:1109–1123. https://doi.org/10.1016/j.cell.2019.03.040.
53. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098–1109. https://doi.org/10.1016/j.cell.2021.01.029.
54. Gregory AC, Zablocki O, Howell A, Bolduc B, Sullivan MB. 2019. The human gut virome database. bioRxiv https://doi.org/10.1101/655910.
55. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, Kyrpides NC. 2021. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol 6:960–970. https://doi.org/10.1038/s41564-021-00928-6.
56. Ackermann HW. 1998. Tailed bacteriophages: the order Caudovirales. Adv Virus Res 51:135–201. https://doi.org/10.1016/s0065-3527(08)60785-x.
57. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 11:119. https://doi.org/10.1186/1471-2105-11-119.
58. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641. https://doi.org/10.1093/nar/27.23.4636.
59. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29: 2607–2618. https://doi.org/10.1093/nar/29.12.2607.
60. Borodovsky M, Lomsadze A. 2014. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. Curr Protoc Microbiol 32:Unit 1E.7. https://doi.org/10.1002/9780471729259.mc01e07s32.
61. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365–8366. https://doi.org/10.1038/srep08365.
62. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, Chlenski P, Conrad N, Dickerman A, Dietrich EM, Gabbard JL, Gerdes S, Guard A, Kenyon RW, Machi D, Mao C, Murphy-Olson D, Nguyen M, Nordberg EK, Olsen GJ, Olson RD, Overbeek JC, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomas C, VanOeffelen M, Vonstein V, Warren AS, Xia F, Xie D, Yoo H, Stevens R. 2020. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res 48: D606–D612. https://doi.org/10.1093/nar/gkz943.
63. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
64. Chirico N, Vianelli A, Belshaw R. 2010. Why genes overlap in viruses. Proc Biol Sci 277:3809–3817. https://doi.org/10.1098/rspb.2010.1052.
65. Wright BW, Ruan J, Molloy MP, Jaschke PR. 2020. Genome modularization reveals overlapped gene topology is necessary for efficient viral reproduction. ACS Synth Biol 9:3079–3090. https://doi.org/10.1021/acssynbio.0c00323.
66. McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. 2019. PHANOTATE: a novel approach to gene identification in phage genomes. Bioinformatics 35:4537–4542. https://doi.org/10.1093/bioinformatics/btz265.
67. Belfort M. 1990. Phage T4 introns: self-splicing and mobility. Annu Rev Genet 24:363–385. https://doi.org/10.1146/annurev.ge.24.120190.002051.
68. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH, Lavigne R, Oksanen HM, Sullivan MB, Jang HB, Simmonds P, Aiewsakun P, Wittman J, Tolstoy I, Brister JR, Kropinski AM, Adriaenssens EM. 2020. Analysis of spounaviruses as a case study for the overdue reclassification of tailed phages. Syst Biol 69: 110–123. https://doi.org/10.1093/sysbio/syz036.
69. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, Antipov D, Pevzner PA, Koonin EV. 2021. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat Commun 12:1044. https://doi.org/10.1038/s41467-021-21350-w.
70. Shapiro JW, Putonti C. 2021. Rephine.r: a pipeline for correcting gene calls and clusters to improve phage pangenomes and phylogenies. PeerJ 9:e11950. https://doi.org/10.7717/peerj.11950.
71. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ, Seed KD, Blekhman R, Aarestrup FM, Thomas BC, Banfield JF. 2019. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol 4:693–700. https://doi.org/10.1038/s41564-018-0338-9.
72. Crisci MA, Chen LX, Devoto AE, Borges AL, Bordin N, Sachdeva R, Tett A, Sharrar AM, Segata N, Debenedetti F, Bailey M, Burt R, Wood RM, Rowden LJ, Corsini PM, van Winden S, Holmes MA, Lei S, Banfield JF, Santini JM. 2021. Closely related Lak megaphages replicate in the microbiomes of diverse animals. iScience 24:102875. https://doi.org/10.1016/j.isci.2021.102875.
73. Dutilh BE, Jurgelenaite R, Szklarczyk R, van Hijum SA, Harhangi HR, Schmid M, de Wild B, Françoijs KJ, Stunnenberg HG, Strous M, Jetten MS, Op den Camp HJ, Huynen MA. 2011. FACIL: fast and accurate genetic code inference and logo. Bioinformatics 27:1929–1933. https://doi.org/10.1093/bioinformatics/btr316.
74. Salisbury A, Tsourkas PK. 2019. A method for improving the accuracy and efficiency of bacteriophage genome annotation. Int J Mol Sci 20: 3391. https://doi.org/10.3390/ijms20143391.
75. Lazeroff M, Ryder G, Harris S, Tsourkas PK. 2021. Phage Commander, a software tool for rapid annotation of bacteriophage genomes using multiple programs. Phage 2:204–213. https://doi.org/10.1089/phage.2020.0044.
76. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinform 10:421. https://doi.org/10.1186/1471-2105-10-421.
77. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
78. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473. https://doi.org/10.1186/s12859-019-3019-7.
79. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. https://doi.org/10.1093/nar/gky1085.
80. Moreno-Gallego JL, Reyes A. 2021. Informative regions in viral genomes. Viruses 13:1164. https://doi.org/10.3390/v13061164.
81. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit M-A, Enault F. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform 3: lqab067. https://doi.org/10.1093/nargab/lqab067.
82. Low SJ, Džunková M, Chaumeil PA, Parks DH, Hugenholtz P. 2019. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. Nat Microbiol 4:1306–1315. https://doi.org/10.1038/s41564-019-0448-z.
83. Andrade-Martínez JS, Moreno-Gallego JL, Reyes A. 2019. Defining a core genome for the Herpesvirales and exploring their evolutionary relationship with the Caudovirales. Sci Rep 9:11342. https://doi.org/10.1038/s41598-019-47742-z.
84. Ecale Zhou CL, Kimbrel J, Edwards R, McNair K, Souza BA, Malfatti S. 2021. MultiPhATE2: code for functional annotation and comparison of phage genomes. G3 (Bethesda) 11:jkab074. https://doi.org/10.1093/g3journal/jkab074.
85. González-Tortuero E, Sutton TD, Velayudhan V, Shkoporov AN, Draper LA, Stockdale SR, Ross RP, Hill C. 2018. VIGA: a sensitive, precise and automatic de novo VIral Genome Annotator. bioRxiv. https://doi.org/10.1101/277509.
86. McNair K, Aziz RK, Pusch GD, Overbeek R, Dutilh BE, Edwards R. 2018. Phage genome annotation using the RAST pipeline, p 231–238. *In* Clokie M, Kropinski A, Lavigne R (ed), Bacteriophages. Humana Press, New York, NY.
87. Harris RS. 2007. Improved pairwise alignment of genomic DNA. PhD thesis. The Pennsylvania State University, Old Main, PA.

88. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935. https://doi.org/10.1093/bioinformatics/btt509.
89. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16. https://doi.org/10.1093/nar/gkh152.
90. Edgar RC. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinform 8:18. https://doi.org/10.1186/1471-2105-8-18.
91. Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580. https://doi.org/10.1093/nar/27.2.573.
92. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res 14:1861–1869. https://doi.org/10.1101/gr.2542904.
93. Abid D, Zhang L. 2018. DeepCapTail: a deep learning framework to predict capsid and tail proteins of phage genomes. bioRxiv. https://doi.org/10.1101/477885.
94. Fang Z, Zhou H. 2021. VirionFinder: identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids. Front Microbiol 12:615711. https://doi.org/10.3389/fmicb.2021.615711.
95. Chu Y, Guo S, Cui D, Fu X, Ma Y. 2021. DeephageTP: a convolutional neural network framework for identifying phage-specific proteins from metagenomic sequencing data. Res Square. https://doi.org/10.21203/rs.3.rs-21641/v2.
96. Cantu VA, Salamon P, Seguritan V, Redfield J, Salamon D, Edwards RA, Segall AM. 2020. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. PLoS Comput Biol 16:e1007845. https://doi.org/10.1371/journal.pcbi.1007845.
97. Chibani CM, Farr A, Klama S, Dietrich S, Liesegang H. 2019. Classifying the unclassified: a phage classification method. Viruses 11:195. https://doi.org/10.3390/v11020195.
98. Mande SS, Mohammed MH, Ghosh TS. 2012. Classification of metagenomic sequences: methods and challenges. Brief Bioinform 13:669–681. https://doi.org/10.1093/bib/bbs054.
99. Iranzo J, Krupovic M, Koonin EV. 2016. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. mBio 7:e00978-16. https://doi.org/10.1128/mBio.00978-16.
100. Turner D, Kropinski AM, Adriaenssens EM. 2021. A roadmap for genome-based phage taxonomy. Viruses 13:506. https://doi.org/10.3390/v13030506.
101. International Committee on Taxonomy of Viruses Executive Committee. 2020. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol 5:668. https://doi.org/10.1038/s41564-020-0709-x.
102. Simmonds P, Aiewsakun P. 2018. Virus classification—where do you draw the line? Arch Virol 163:2037–2046. https://doi.org/10.1007/s00705-018-3938-z.
103. Aiewsakun P, Simmonds P. 2018. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome 6:38. https://doi.org/10.1186/s40168-018-0422-7.
104. Moraru C, Varsani A, Kropinski AM. 2020. VIRIDIC—a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. Viruses 12:1268. https://doi.org/10.3390/v12111268.
105. Meier-Kolthoff JP, Göker M. 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 33:3396–3404. https://doi.org/10.1093/bioinformatics/btx440.
106. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. 2017. ViPTree: the viral proteomic tree server. Bioinformatics 33:2379–2380. https://doi.org/10.1093/bioinformatics/btx157.
107. Dougan TJ, Quake SR. 2019. Viral taxonomy derived from evolutionary genome relationships. PLoS One 14:e0220440. https://doi.org/10.1371/journal.pone.0220440.
108. Pons JC, Paez-Espino D, Riera G, Ivanova N, Kyrpides NC, Llabrés M. 2021. VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. Bioinformatics 37:1805–1813. https://doi.org/10.1093/bioinformatics/btab026.
109. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Bork P, Acinas SG, Wincker P, Sullivan MB, Tara Oceans Coordinators. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689–693. https://doi.org/10.1038/nature19366.
110. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639. https://doi.org/10.1038/s41587-019-0100-8.
111. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev 40:258–272. https://doi.org/10.1093/femsre/fuv048.
112. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. 2017. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res 45:39–53. https://doi.org/10.1093/nar/gkw1002.
113. Coutinho FH, Zaragoza-Solas A, López-Pérez M, Barylski J, Zielezinski A, Dutilh BE, Edwards RA, Rodriguez-Valera F. 2020. RaFAH: A superior method for virus-host prediction. bioRxiv. https://doi.org/10.1101/2020.09.25.313155.
114. Galiez C, Siebert M, Enault F, Vincent J, Söding J. 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics 33:3113–3114. https://doi.org/10.1093/bioinformatics/btx383.
115. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, Larsen MV. 2016. HostPhinder: a phage host prediction tool. Viruses 8:116. https://doi.org/10.3390/v8050116.
116. Gałan W, Bąk M, Jakubowska M. 2019. Host taxon predictor-a tool for predicting taxon of the host of a newly discovered virus. Sci Rep 9:3436. https://doi.org/10.1038/s41598-019-39847-2.
117. Liu D, Ma Y, Jiang X, He T. 2019. Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion. BMC Bioinform 20(Suppl 16):594. https://doi.org/10.1186/s12859-019-3082-0.
118. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, Braun J, Sun F, Ahlgren NA. 2020. A network-based integrated framework for predicting virus–prokaryote interactions. NAR Genom Bioinform 2:lqaa044. https://doi.org/10.1093/nargab/lqaa044.
119. Zhang R, Mirdita M, Levy Karin E, Norroy C, Galiez C, Söding J. 2021. SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. Bioinformatics 37:3364–3366. https://doi.org/10.1093/bioinformatics/btab222.
120. Dion MB, Plante PL, Zufferey E, Shah SA, Corbeil J, Moineau S. 2021. Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. Nucleic Acids Res 49:3127–3138. https://doi.org/10.1093/nar/gkab133.
121. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM. 2016. CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC Genom 17:356. https://doi.org/10.1186/s12864-016-2627-0.
122. Laanto E, Hoikkala V, Ravantti J, Sundberg LR. 2017. Long-term genomic coevolution of host-parasite interaction in the natural environment. Nat Commun 8:111. https://doi.org/10.1038/s41467-017-00158-7.
123. Zhang F, Zhou F, Gan R, Ren C, Jia Y, Yu L, Huang Z. 2020. PHISDetector: a tool to detect diverse in silico phage-host interaction signals for virome studies. bioRxiv. https://doi.org/10.1101/661074.
124. Amgarten D, Iha BK, Piroupo CM, da Silva AM, Setubal JC. 2020. vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. bioRxiv. https://doi.org/10.1101/2020.12.06.413476.
125. Zhang M, Yang L, Ren J, Ahlgren NA, Fuhrman JA, Sun F. 2017. Prediction of virus-host infectious association by supervised learning methods. BMC Bioinform 18:143–154.
126. Mirzaei MK, Maurice CF. 2017. Ménage à trois in the human gut: interactions between host, bacteria and phages. Nat Rev Microbiol 15:397–408. https://doi.org/10.1038/nrmicro.2017.30.
127. McNair K, Bailey BA, Edwards RA. 2012. PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28:614–618. https://doi.org/10.1093/bioinformatics/bts014.
128. Hockenberry AJ, Wilke CO. 2021. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ 9:e11396. https://doi.org/10.7717/peerj.11396.
129. Paul JH. 1999. Microbial gene transfer: an ecological perspective. J Mol Microbiol Biotechnol 1:45–50.
130. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 110:12450–12455. https://doi.org/10.1073/pnas.1300833110.

131. Buckling A, Rainey PB. 2002. Antagonistic coevolution between a bacterium and a bacteriophage. Proc R Soc Lond B 269:931–936. https://doi.org/10.1098/rspb.2001.1945.

132. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A, Guerin E, Velayudhan V, Ross RP, Hill C. 2019. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe 26:527–541. https://doi.org/10.1016/j.chom.2019.09.009.

133. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. 2020. Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat Microbiol 5: 265–271. https://doi.org/10.1038/s41564-019-0628-x.

134. De Sordi L, Lourenço M, Debarbieux L. 2019. "I will survive": a tale of bacteriophage-bacteria coevolution in the gut. Gut Microbes 10:92–99. https://doi.org/10.1080/19490976.2018.1474322.

135. Enav H, Kirzner S, Lindell D, Mandel-Gutfreund Y, Béjà O. 2018. Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. Nat Commun 9:4698. https://doi.org/10.1038/s41467-018-07164-3.

136. Coutinho FH, Rosselli R, Rodríguez-Valera F. 2019. Trends of microdiversity reveal depth-dependent evolutionary strategies of viruses in the Mediterranean. mSystems 4:e00554-19. https://doi.org/10.1128/mSystems.00554-19.

137. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. https://doi.org/10.1038/ng.806.

138. Chen LX, Méheust R, Crits-Christoph A, McMahon KD, Nelson TC, Slater GF, Warren LA, Banfield JF. 2020. Large freshwater phages with the potential to augment aerobic methane oxidation. Nat Microbiol 5: 1504–1515. https://doi.org/10.1038/s41564-020-0779-9.

139. Siranosian BA, Tamburini FB, Sherlock G, Bhatt AS. 2020. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat Commun 11:280. https://doi.org/10.1038/s41467-019-14103-3.

140. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P. 2013. Genomic variation landscape of the human gut microbiome. Nature 493:45–50. https://doi.org/10.1038/nature11711.

141. Cosgun E, Oh M. 2020. Exploring the consistency of the quality scores with machine learning for next-generation sequencing experiments. Biomed Res Int 2020:8531502. https://doi.org/10.1155/2020/8531502.

142. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E, Langley CH. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol 5:e310. https://doi.org/10.1371/journal.pbio.0050310.

143. Gregory AC, Gerhardt K, Zhong ZP, Bolduc B, Temperton B, Konstantinidis KT, Sullivan MB. 2020. MetaPop: A pipeline for macro-and micro-diversity analyses and visualization of microbial and viral metagenome-derived populations. bioRxiv. https://doi.org/10.1101/2020.11.01.363960.

144. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. 2021. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol 39:727–710. https://doi.org/10.1038/s41587-020-00797-0.

145. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J Veg Sci 14:927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x.

146. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. https://doi.org/10.1371/journal.pone.0061217.

147. Szekely AJ, Breitbart M. 2016. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. FEMS Microbiol Lett 363:fnw027. https://doi.org/10.1093/femsle/fnw027.

148. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. 2018. RNA phage biology in a metagenomic era. Viruses 10:386. https://doi.org/10.3390/v10070386.

149. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.

150. UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049.

151. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. Nucleic Acids Res 49: D412–D419. https://doi.org/10.1093/nar/gkaa913.

152. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48: D265–D268. https://doi.org/10.1093/nar/gkz991.

153. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. 2014. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42:D310–D314. https://doi.org/10.1093/nar/gkt1242.

154. Andreeva A, Kulesha E, Gough J, Murzin AG. 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 48:D376–D382. https://doi.org/10.1093/nar/gkz1064.

155. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. 2017. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat Protoc 12:1673–1682. https://doi.org/10.1038/nprot.2017.063.